Taking account

As a member of the SIGCOMM TPC this year, I recently had a chance to read over thirty submissions by the best and brightest in the field. Imagine my surprise to find that all but one of them had significant flaws in statistical analysis. These flaws were severe enough that I suspect that in other disciplines the papers would have been summarily rejected. Yet, a few of them not only were accepted but also are likely to become role models for the next generation of students. For, though statistically flawed, they were not far from common practice. Moreover, it was felt that the papers should not be unfairly punished.

In this editorial, I will focus on why statistical analysis matters, three of the most common statistical errors I saw, why I think we have had, as a field, a rather relaxed approach to statistical analysis, and what we can do about it.

To begin with, it is an unfortunate fact that only the most trivial situations allow a comprehensive exploration of the underlying parameter space: simulations and measurements alike allow us to explore only a small portion of the space. One role of statistical analysis is guide the selection of the parameter space to be explored using techniques from experimental design.

A second role of statistical analysis is to allow a researcher to draw cautious and justifiable conclusions from a mass of numbers. Over a hundred years of work has created tried-and-tested techniques that allow researchers to compensate for unavoidable measurement errors, and to infer with high probability that the improvement seen due a particular algorithm or system is *significant*, rather than due to mere luck. Without statistical analysis, one is on thin ice.

These two roles of statistical analysis make it an essential underpinning for experimental research, especially in the area of experimental design, measurement, and performance analysis.

Unfortunately, despite its importance, papers in our field--both submissions to SIGCOMM and published papers in similar top-tier conferences--suffer from severe statistical errors. The three most common errors are: (1) confusing a sample with a population and, as a corollary, not specifying the underlying population (2) not presenting confidence intervals for sample statistics and (3) incorrect hypothesis testing.

Most authors did not seem to realize that their measurements represented a sample from a much larger underlying population. Consider the measured throughput during ten runs between two nodes using a particular wireless protocol. These values are a sample of the population of node-to-node throughputs obtained using all possible uses of the protocol under all conceivable circumstances. Wireless performance may, however, vary widely depending on the RF environment. Therefore, the sample can be considered to be

representative only if every likely circumstance had a chance of being represented. Authors need to strongly argue that the sample measurements are chosen in a way that sufficiently covers the underlying parameter space. Otherwise, the sample represents nothing more than itself! Yet, this obvious criterion for scientific validity is rarely discussed in most papers.

Given that a sample is not the population, it is imperative that the statistics of a sample be presented along with a confidence interval in which, with high confidence, the population parameters lie. These are the familiar error bars in a typical graph or histogram. Lacking error bars, we cannot interpret the characteristics of the population with any precision; we can only draw conclusions about the sample, which is necessarily limited. To my surprise, only one paper I read had error bars! This is a serious flaw in analysis.

Finally, it is axiomatic in statistical analysis that a hypothesis cannot be proved; it can only rejected or not rejected. Hypothesis testing requires carefully framing a null hypothesis, and then using standard statistical analysis to either reject or not reject it. I agree that in many cases the null hypothesis is obvious and need not be formally stated. Yet, *formulating* the underlying hypothesis, and then being cautious when interpreting results is both essential and sorely lacking.

Why do papers in our field lack statistical rigor? I suspect that part of the reason is that we teach statistics early in the academic curriculum. Students who learn statistical inference and hypothesis testing as a chore in a freshman class have all but forgotten it by the time they are writing papers. In my own case, I am embarrassed to admit that I never thoroughly understood these techniques until I wrote a chapter on statistical techniques for the second edition of my book. I suspect that many of my colleagues are in the same boat. Unfortunately, this makes our weakness self-perpetuating. Having forgotten statistical analysis, we are neither in a position to carry it out properly, nor do we insist upon it during peer review. Thus, we stumble from one flawed paper to the next, continuing the cycle.

What can we do about this? I suggest that all graduate students be required to take a course on statistical analysis. This need not be a formal course, but could be taken online or using distance education. The concepts are well known and the techniques are thoroughly explained in numerous textbooks. We just need to buy into the agenda. Second, I think that we need to raise the bar during paper evaluation. Poor statistical analysis should be pointed out and should form one criterion for paper rejection. For papers that are novel and thorough, but have poor statistical analysis, we should insist that these issues be rectified during shepherding. Finally, we need to educate the educators. Perhaps SIGCOMM can sponsor online or offline tutorials where researchers can quickly come up to speed in statistical analysis.

If we do this, and I think we should, then we can raise the scientific merit of our discipline, and, more importantly, not be misled into accepting incorrect results due to flaws in statistical analysis.